TRABAJO DE FIN DE MÁSTER

Máster en Estadística e Investigación Operativa

Título: Principal component analysis of bi-allelic genetic marker data

Autor: Genny Paola Díaz Rodríguez

Director: Jan Graffelman

Departamento: EIO - Departament d'Estadística i Investigació Operativa

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya

Treball De Fi De Màster

Principal component analysis of bi-allelic genetic marker data

Genny Paola Díaz Rodríguez

Director: Jan Graffelman

EIO - Departament d'Estadística i Investigació Operativa

Dedicatoria a:

Dios, por permitirme tener salud y la fortaleza necesaria para no desfallecer y culminar esta memoria. A mi madre Luz Nelly, por darme la vida, por creer en mí y por el amor que me brinda siempre. A mi esposo Andrés Felipe, mis hijos Gabriela y Juan Martín por ser mi motivación para seguir adelante, por cada palabra de inspiración, por la compañía y por el apoyo incondicional que me brindaron. A mi director del proyecto Jan Graffelman, por la paciencia, por compartir su conocimiento conmigo y por encaminarme para que fuera posible la realización de esta memoria. A todos las personas de la UPC que intervinieron para finalizar mi proyecto desde la distancia.

Resumen

En los estudios del genoma humano se producen grandes bases de datos de polimorfismos de un solo nucleótido (los SNPs). El análisis de estas bases de datos presenta muchos retos estadísticos. En esta memoria se plantea el análisis de este tipo de datos mediante una herramienta clásica del análisis multivariante: el análisis de componentes principales (ACP). Dentro este marco se presentan una introducción a los datos genéticos modernos y un resumen de la teoría del ACP.

Los polimorfismos genéticos son variables categóricas, pero mediante el conteo de sus alelos, estas se pueden convertir en variables cuantitativas, que solo toman los valores 0, 1 y 2. Esta conversión abre la vía para la aplicación del ACP a los datos genéticos. En la memoria se estudia el efecto de la codificación empleada sobre el ACP, y se demuestra la invarianza de los resultados respecto al alelo contado.

Se presentan aplicaciones del ACP a los polimorfismos de dos genes relacionados con la musculatura humana, los genes ACTN3 (alpha-actinin 3) y AKT1 (serine-threonine protein kinase), utilizando la base de datos FAMuSS, que recoge información de 225 SNPs para una muestra de 1397 personas. Se consideran tanto el ACP basado en correlaciones como el ACP basado en covarianzas, teniendo en cuenta la población de procedencia de los individuos.

Se obtienen representaciones gráficas (biplots) con componentes interpretables y que permiten identificar grupos de SNPs estrechamente relacionados, demostrando así la utilidad del método para el análisis de los polimorfismos genéticos.

Palabras clave: Datos genéticos. Alelos. Polimorfismos de un solo nucleótido (SNPs). Análisis de Componentes Principales (ACP). Biplot.
MSC2000: 62-07. 62H25. 62H35.

Abstract

The study of the human genome has produced large databases of single nucleotide polymorphisms (SNPs). The analysis of these databases presents many statistical challenges. This project consider to analyze SNP data with a classical tool from the field of multivariate analysis: principal component analysis (PCA). Within this context, a brief introduction to modern genetic data and a theoretical summary of PCA are presented.

Genetic polymorphisms are categorical variables, but by counting alleles they can be converted into quantitative variables, taking on only the values 0, 1 or 2. This conversion makes it possible to analyze SNP data by PCA. It is shown that PCA results are invariant with respect to the chosen coding for the conversion, and that it makes no difference which of the two alleles is counted.

Applications of PCA are presented using polymorphisms of two genes related to human muscle activity, the genes ACTN3 (alpha-actinin 3) and AKT1 (serine-threonine protein kinase), using the database FAMuSS, which contains information on 225 SNPs for a sample of 1397 individuals. Both correlation-based and covariance-based PCA are considered, and the analysis is stratified with respect to the human genetic background of the individuals.

The analysis leads to graphical representations (biplots) with interpretable components that allow one to identify groups of closely correlated polymorphisms, so demonstrating the usefulness of PCA for the analysis of genetic polymorphisms.

Keywords: Genetic data. Alleles. Single nucleotide polymorphisms (SNPs). Principal Component Analysis (PCA). Biplot.
MSC2000: 62-07. 62H25. 62H35.

Índice

Índice de tablas

Índice de figuras

Capítulo I. Introducción

1) Preliminares

En la actualidad es de interés para la investigación realizar estudios de asociación del genoma humano completo. Estos análisis se realizan para encontrar variaciones genéticas a lo largo del genoma con el propósito en encontrar asociaciones a un rasgo observable. Es por ello que este tipo de investigaciones se realizan principalmente a las asociaciones entre los polimorfismos de un solo nucleótido y rasgos como enfermedades principales.

Para proceder a realizar estos estudios, se debe contar con datos genéticos procedentes de varios individuos, de tal modo que a partir de las secuenciaciones de los genomas se puedan identificar genes ligados a enfermedades. Los Polimorfismos de un Solo Nucleótido (SNPs) son esenciales para estas investigaciones debido a que se puede contrastar cómo se produce la aparición de ellos en alguna secuencia del genoma siempre que aparece el mismo fenotipo, con lo cual se puede establecer que el cambio presentado a nivel genético corresponde a un rasgo significativo y que principalmente se asocia a enfermedades.

Este tipo de investigaciones en humanos ha permitido descubrir patrones que indican que la variación de ciertos genes está ligada por ejemplo a enfermedades como la degeneración muscular asociada a la edad y la diabetes. Existen cerca de 600 estudios de asociación del genoma humano completo a partir del análisis realizado a miles de personas a las que se les ha encontrado aproximadamente 800 SNPs asociados a alrededor de 150 rasgos y enfermedades (Feero, 2010; Pearson, 2008).

En los estudios genéticos modernos se recoge mucha información y estos estudios terminan siendo del ámbito multivariado, por tal razón en esta memoria se consideran métodos de la naturaleza del análisis multivariante para estudiar los datos. El presente trabajo tiene como objetivo investigar la utilidad del Análisis de Componentes Principales (ACP) como método para analizar datos de polimorfismos genéticos. Este método puede ser de gran importancia en estudios de datos genéticos debido a que si la base de datos tiene una dimensión muy grande se puede realizar una reducción del número de variables a partir de la combinación lineal que resulte de las mismas, conservando un porcentaje significativo de la información original.

En la presente memoria se utiliza la base de datos FAMuSS (**F**unctional SNPs **A**ssociated with **Mu**scle **S**ize and **S**trength) creada por un instituto financiado de salud con el propósito de identificar los SNPs en proteínas musculares y determinar cuáles de ellos contribuyen en el tamaño del musculo, su flexor, fuerza y respuesta después de 12 semanas de entrenamiento con ejercicios de resistencia.

La presente memoria consta de 5 capítulos. En el primer capítulo se encuentra una introducción a los datos genéticos y la explicación específica de la base de datos FAMuSS. En el segundo capítulo se repasa la teoría del análisis de componentes principales y la elaboración de biplots. En el tercer capítulo trata la aplicación del ACP a los datos genéticos, donde se usan los polimorfismos de la base de datos FAMuSS. En

el capítulo 4 se presentan las principales conclusiones del estudio. La bibliografía (capítulo 5) completa esta memoria.

1.1) Introducción a los datos genéticos.

La vida se presenta de formas variadas, la célula como componente principal, conformando desde organismos unicelulares hasta complejos sistemas orgánicos de especies superiores. Su componente más importante se encuentra en el núcleo y contiene toda la información necesaria para formar el individuo, el Acido Desoxirribonucleico o ADN.

El paradigma actual considera al ADN como la molécula de la vida, está presente en todos los organismos vivos, compuesta por los mismos elementos, variando solo en su disposición y cantidad para determinar rasgos tan evidentes como las diferencias entre especies hasta particularidades en individuos de una misma especie.

El ADN se compone de nucleótidos, estos a su vez son conformados por una de cada una de las siguientes moléculas: Grupos fosfato, azúcar desoxirribosa (en conjunto forman la parte estructural), y bases nitrogenadas (Adenina (A), Guanina (G), Timina (T) y Citosina (C)), la unión lineal de nucleótidos forman una hebra de ADN.

El ADN está compuesto por dos cadenas complementarias y anti paralelas. La complementariedad hace referencia a que la secuencia de bases de una cadena está relacionada en un estricto orden de apareamiento con la secuencia de bases de la cadena enfrentada, Timina se relacionará con Adenina y Guanina se relacionará con Citosina. Las cadenas son anti paralelas porque la dirección de los nucleótidos de una hebra (por ejemplo: 5’ – 3’) es contraria a la de la hebra complementaria (3’ – 5’).

La secuencia de ADN encontrada en una célula humana consta alrededor de 3000 millones de pares de bases, y dentro de ella no todos los segmentos codificaran para un gen.

Todas las posibles combinaciones de la secuencia que codifican un gen se denominan alelos. Estas diferencias alélicas crean la base para la expresión fenotípica de los organismos con sus diferencias y similitudes, son varios los factores que pueden alterar el orden de una secuencia, desde la misma recombinación genética durante el proceso de fecundación, pasando por factores físico químicos internos y externos, hasta eventos espontáneos. Las alteraciones en la secuencia pueden eliminar, intercambiar o adicionar nucleótidos (incluso segmentos completos) a la secuencia. El resultado final de estos procesos puede, o no, alterar la expresión de un gen que podría tener repercusiones a nivel biológico en el individuo, se ha postulado que estos procesos con adición o eliminación de características son la base del proceso evolutivo.

Existen diversas variaciones genéticas en el genoma humano. Los polimorfismos de un solo nucleótido (SNPs) son los que representan las variaciones más comunes de una sola base en el ADN. Los SNPs se encuentran en cualquier parte de la estructura del genoma o de los genes. En la actualidad son tema de estudio de muchos científicos que buscan correlacionarlos con enfermedades, respuestas a medicamentos y con otros fenotipos. Para que estas variaciones sean consideradas SNP deben presentarse en por lo

menos el 1% de la población, en caso contrario se considera como caso especial o mutación puntual.

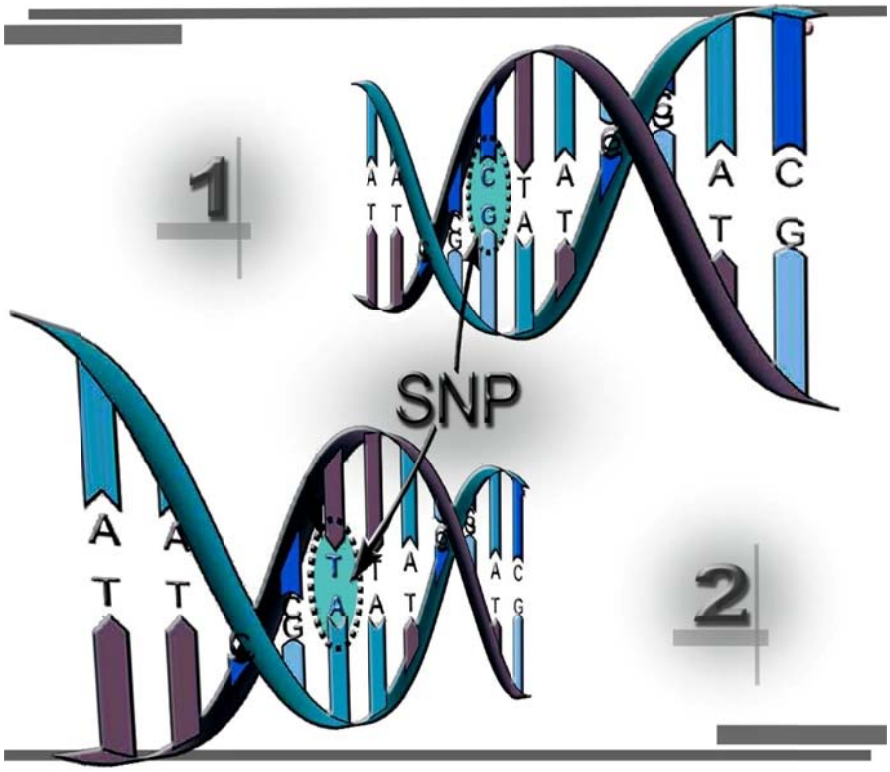

Figura 1.1Polimorfismo de un Solo Nucleótido SNPs

En la figura 1.1 se muestra un en ejemplo de un SNP, representando los dos cromosomas de un individuo. El cambio de bases en la secuencia es lo que indica que existe un polimorfismo de un solo nucleótido, en este caso un polimorfismo A/G.

Se dice que un individuo es homocigoto, si son idénticos el par de alelos homólogos, mientras que si son distintos se denomina heterocigoto.

De acuerdo al número de alelos, así mismo un polimorfismo puede tener cierto número de genotipos. Por ejemplo, un SNP bialelico tiene 3 genotipos y para un trialelico hay 6 genotipos. En general un SNP con k alelos tiene $0{,}5\ k\ (k\ +\ 1)$ genotipos.

Estadísticamente los SNPs se consideran variables de tipo categóricas, esto se debe a que sus respuestas vienen dadas a partir de la combinación de las bases nitrogenadas que las componen. Las posibles categorías que se den varían de acuerdo a cada marcador y el número de alelos que éste contenga. Generalmente, lo más común es encontrar SNPs para los cuales existen 3 genotipos diferentes. Por ejemplo, si el marcador consta de los alelos A y T, entonces los 3 genotipos posibles serán AA, AT y TT, es decir, todas las posibles combinaciones que puedan resultar a partir de los dos alelos. Para realizar el análisis de este tipo de variables es necesario realizar una codificación a partir del cálculo del alelo menor o el alelo mayor, esto se hace en base a las frecuencias alélicas relativas del marcador como se verá más adelante.

1.2) Base de datos FAMuSS.

Para analizar la utilidad del ACP para este tipo de datos se emplea, a lo largo de esta memoria, una base de datos del estudio FAMuSS (**F**unctional SNPs **A**ssociated with **Mu**scle **S**ize and **S**trength) en el que se toman polimorfismos de un solo nucleótido (o por sus siglas en ingles SNP) asociados al crecimiento del musculo y la fuerza humana. En dicho estudio emplean 1397 individuos sanos, aproximadamente 59% mujeres y 41% hombres principalmente de raza caucásica. Hay un total de 347 variables donde 225 son SNPs (Foulkes, 2009; Thompson, 2004).

Para el estudio estadístico y debido a que las variables genéticas son de tipo categórico, se realiza una codificación que consiste en la asignación numérica de las categorías de cada SNP de acuerdo a la frecuencia del alelo menor.

Se tomara el SNP actn3_r577x como ejemplo para ilustrar el proceso de identificación y codificación del alelo menor. La tabla de frecuencias para esta variable se muestra a continuación

actn3_r577x	**Frecuencia Absoluta**	**Frecuencia Relativa**
CC	216	0,2939
CT	318	0,4327
TT	201	0,2735
NA	662	---
Total	1397	

Tabla 1.1 Ejemplo de frecuencias genotípicas

De la tabla anterior se puede identificar que la frecuencia del genotipo CC es 216, la frecuencia del genotipo CT es 318, la frecuencia del genotipo TT es 201 y que hay un total de 662 de observaciones faltantes identificadas como NA. Por razones de simplicidad, estos valores faltantes se omiten en el cálculo de las frecuencias, por lo cual el total para este cálculo es de 735 observaciones.

Las frecuencias genotípicas están dadas por:

$$f_{CC} = \frac{216}{735} = 0{,}2939$$

$$f_{CT} = \frac{318}{735} = 0{,}4327$$

$$f_{TT} = \frac{201}{735} = 0{,}2735$$

Las frecuencias alélicas para C y T están dadas por:

$$f_C = \frac{2f_{CC} + f_{CT}}{2} = 0{,}5102$$
$$f_T = \frac{2f_{TT} + f_{CT}}{2} = 0{,}4898$$

Por lo tanto T es el alelo menor para SNP actn3_r577x. Una vez identificado el alelo menor, se procede a la codificación de esta variable asignando el valor de 2 al genotipo TT por ser el del menor alelo, 1 a CT y 0 a CC.

Una manera alternativa para el cálculo de frecuencias alélicas y genotípicas es usar las funciones **genotype()** y **summary()** del paquete **genetics()** en R (Gregory Warnes et al., 2013), donde se obtiene:

```
Number of samples typed: 735 (52.6%)


Allele Frequency: (2 alleles)
   Count Proportion
C    750       0.51
T    720       0.49
NA  1324         NA

Genotype Frequency:
    Count Proportion
C/C   216       0.29
C/T   318       0.43
T/T   201       0.27
NA    662         NA

Heterozygosity (Hu)  = 0.500132
Poly. Inf. Content   = 0.3748959
```

Lo cual confirma que para el locus SNP actn3_r577x el alelo menor es T y C es el alelo mayor. Decir que T es el alelo menor significa que TT es el homocigoto con la frecuencia más baja en actn3_r577x. Mientras que debido a que C es el alelo más común, se denominará CC el homocigoto de tipo natural.

Capítulo 2. Análisis de Componentes Principales

2) Teoría básica

2.1) Teoría del Análisis de Componentes Principales

El análisis de componentes principales (ACP), es un método estadístico utilizado con el fin de realizar una reducción significativa del número de variables con la menor pérdida de información posible. Las componentes principales creadas son independientes entre sí y son una combinación lineal de las variables originales (Johnson & Wichern, 2002; Jolliffe, 1986).

Dada una matriz X de dimensión $n \times p$, donde n corresponde al número de datos y p al número de variables, se pretende crear una nueva matriz de componentes principales de m variables independientes entre sí, con $m < p$ y que sean combinaciones lineales de las p variables originales y que describa en un gran porcentaje la información original. Estas componentes principales se tendrán en cuenta por el orden de importancia de la variabilidad recogida de los datos.

2.2) Cálculo de las Componentes Principales

Sean $X_j = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}$ con $j = 1,2, \dots, p$; una serie de variables con información de n individuos. Se calcula a partir de estos datos un nuevo conjunto de variables $\mathrm{Z}_\mathrm{j} = (\mathrm{Z}_1, \mathrm{Z}_2, \dots, \mathrm{Z}_\mathrm{p})$ independientes entre sí y con varianzas decrecientes progresivamente que son resultado de combinaciones lineales de las variables originales, es decir:

$$\mathrm{Z}_\mathrm{p} = \mathrm{a}_{\mathrm{p}1}X_1 + \mathrm{a}_{\mathrm{p}2}X_2 + \cdots + \mathrm{a}_{\mathrm{pp}}X_p$$

Donde a_j es un vector de escalares entre -1 y 1, y

$$\mathrm{a}_{i1}^2 + \mathrm{a}_{i2}^2 + \cdots + \mathrm{a}_{ip}^2 = 1$$

La primera componente principal se calcula escogiendo a_1 de tal manera que Z_1 tenga la mayor varianza posible, sujeto a la restricción $\mathrm{a}_1'\mathrm{a}_1 = 1$. La segunda componente se calcula seleccionando a_2 de manera que Z_2 no esté correlacionada con Z_1. De esta forma se calculan el resto de componentes principales teniendo en cuenta que las variables obtenidas tengan cada vez menor varianza y que las variables $\mathrm{Z}_1, \mathrm{Z}_2, \dots, \mathrm{Z}_\mathrm{p}$ no estén correlacionadas.

Sea X un vector aleatorio de dimensión $p \times 1$ con matriz de covarianza Σ y Z_1 la combinación lineal $\mathrm{a}_1'X$, se desea elegir a a_1 de tal manera que se maximice la varianza de Y_1 sujeta a la restricción $\mathrm{a}_1'\mathrm{a}_1 = 1$, es decir:

$$Var(Z_1) = \mathrm{a}_1'XX'\mathrm{a}_1 = \mathrm{a}_1'\Sigma\mathrm{a}_1$$

Usualmente el método que se utiliza para maximizar una función de varias variables que se encuentra sujeta a restricciones es el método de los multiplicadores de Lagrange.

El vector desconocido que nos lleva a obtener la combinación lineal óptima es a_1, por lo tanto se debe construir una función de multiplicadores de Lagrange $\mathcal{L}$.

$$\mathcal{L}(a_1, \lambda_1) = a_1' \Sigma a_1 - \lambda_1 (a_1' a_1 - 1)$$

Derivando con respecto a a_1 y a λ_1, e igualando a cero:

$$\frac{\partial \mathcal{L}}{\partial a_1} = 2\Sigma a_1 - 2\lambda_1 a_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - a_1' a_1 = 0$$

$$\Sigma a_1 = \lambda_1 a_1 \qquad (1)$$

Por lo tanto λ_1 y a_1 son valores y vectores propios de Σ.

La matriz de varianzas covarianzas Σ es definida positiva y de orden p, por tal razón tendrá p valores propios diferentes $\lambda_1, \lambda_2, \dots, \lambda_p$ tales que $\lambda_1 > \lambda_2 > \cdots > \lambda_p$

Si se multiplica (1) por a_1' entonces,

$$Var(Z_1) = a_1' \Sigma a_1 = a_1' a_1 \lambda_1 = \lambda_1$$

Donde λ_1 es la varianza de la primera componente principal y el primer valor propio de Σ con a_1 como su vector propio asociado.

Análogamente se calcula la segunda componente $Z_2 = a_2' X$, pero ahora se requiere que Z_2 no esté correlacionada Z_1, es decir, que $Cov(Z_2, Z_1) = 0$ y por lo tanto,

$$Cov(Z_2, Z_1) = Cov(a_2' X, a_1' X)$$

$$= a_2' \, E[(X - \mu)(X - \mu)'] a_1$$

$$= a_2' \Sigma \, a_1$$

Es decir, para que Z_2 y Z_1 sean incorreladas se necesita que

$$a_2' \Sigma \, a_1 = 0$$

De (1) se sabe que $\Sigma a_1 = \lambda_1 a_1$, multiplicando por a_2', entonces,

$$a_2' \Sigma a_1 = a_2' \lambda_1 a_1 = \lambda_1 a_2' a_1 = 0$$

Por lo que si $\lambda_1 \neq 0$, los vectores a_1 y a_2 serán ortogonales.

La varianza de Z_2 es ahora:

$$Var(Z_2) = a_2' X X' a_2 = a_2' \Sigma a_2$$

Se desea maximizar la función $\mathrm{a}_2'\Sigma\mathrm{a}_2$, sujeta a las restricciones:

$$\mathrm{a}_2'\mathrm{a}_2 = 1 \quad \text{y} \quad \mathrm{a}_1'XX'\mathrm{a}_2 = \mathrm{a}_1'\Sigma\mathrm{a}_2 = 0$$

Aplicando nuevamente el método de los multiplicadores de Lagrange, se tiene:

$$\mathcal{L}(\mathrm{a}_2, \lambda_2, \mu) = \mathrm{a}_2'\Sigma\mathrm{a}_2 - \lambda_2(\mathrm{a}_2'\mathrm{a}_2 - 1) - \mu\mathrm{a}_1'\Sigma\mathrm{a}_2$$

Se deriva y se iguala a cero y por lo tanto:

$$\frac{\partial\mathcal{L}}{\partial\mathrm{a}_2} = 2\Sigma\mathrm{a}_2 - 2\lambda_2\mathrm{a}_2 - \mu\Sigma\mathrm{a}_1 = 0, \qquad (2)$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_2} = 1 - \mathrm{a}_2'\mathrm{a}_2 = 0$$

$$\frac{\partial\mathcal{L}}{\partial\mu} = \mathrm{a}_1'\Sigma\mathrm{a}_2 = 0$$

Si se multiplica la ecuación (2) por a_1', entonces

$$2\mathrm{a}_1'\Sigma\mathrm{a}_2 - \lambda\mu = 0$$

Esto se debe a que:

$$\mathrm{a}_1'\mathrm{a}_2 = \mathrm{a}_2'\mathrm{a}_1 = 0 \text{ y } \mathrm{a}_1'\mathrm{a}_1 = 1$$

De modo que:

$$\mu = 2\mathrm{a}_1'\Sigma\mathrm{a}_2 = 2\mathrm{a}_2'\Sigma\mathrm{a}_1 = 0, \text{ pues } Cov(Z_2, Z_1) = 0.$$

Por lo anterior,

$$\frac{\partial\mathcal{L}}{\partial\mathrm{a}_2} = 2\Sigma\mathrm{a}_2 - 2\lambda_2\mathrm{a}_2 - \mu\Sigma\mathrm{a}_1 = 2\Sigma\mathrm{a}_2 - 2\lambda_2\mathrm{a}_2$$

$$\Sigma\mathrm{a}_2 = \lambda_2\mathrm{a}_2$$

Donde λ_2 es la varianza de la segunda componente principal y el segundo valor propio de Σ con a_2 como su vector propio asociado.

Procediendo de la misma manera, la componente principal j-ésima se obtiene usando el vector propio j de la matriz de varianzas covarianzas Σ y su varianza esta dada por el valor propio j.

Entonces los Z_j componentes calculados se pueden expresar como el producto de una matriz formada por los vectores propios y el vector aleatorio X, es decir,

$$Z = AX$$

Donde,

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_P \end{pmatrix}, \quad A = \begin{pmatrix} \mathrm{a}_{11} & \mathrm{a}_{12} & \cdots & \mathrm{a}_{1p} \\ \mathrm{a}_{21} & \mathrm{a}_{22} & \cdots & \mathrm{a}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{a}_{p1} & \mathrm{a}_{p2} & \cdots & \mathrm{a}_{pp} \end{pmatrix}, \quad X_j = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{pmatrix}$$

Debido a que:

$$Var(Z_j) = \lambda_j$$

Y dado que Z_j se han construido como variables incorreladas, entonces se define la matriz de covarianzas como:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

$$\Lambda = Var(\mathrm{Z}) = \ A' Var(X)\ A = A' \Sigma\ \mathrm{A}$$

Todos los coeficientes y valores propios pueden ser obtenidos a partir de la descomposición espectral de la matriz de varianzas covarianzas:

$$\Sigma = A \Lambda A'$$

2.3) Porcentajes de Variabilidad

Se definió anteriormente a los valores propios λ_j asociados a cada vector propio a_j de la matriz de varianzas covarianzas Σ, como la varianza de cada componente principal Z_j. Por lo tanto al realizar la suma de todos los valores propios, se obtendrá el valor de la variabilidad total de las componentes principales, es decir,

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(\ Z_j),$$

Pero dado a que Λ es una matriz diagonal creada a partir de los valores propios, entonces,

$$\sum_{j=1}^{p} \lambda_j = traza(\Lambda)$$

Y por propiedades de la traza,

$$traza(\Lambda) = traza(\mathrm{A}\ \Sigma\ A') = traza(\Sigma\ A' A) = traza(\Sigma)$$

Pues $A' A = A\ A' = I$ por ser A ortogonal y por lo tanto:

$$traza(\Lambda) = traza(\Sigma) = \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{Var}(X_j)$$

Lo que significa que la suma de las varianzas de las componentes principales coincide con la suma de las varianzas de las variables originales y con lo cual,

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{\sum_{j=1}^{p} \mathrm{Var}(X_j)}$$

Representa la fracción de la varianza total representada por componente j.

Si se desea conocer el porcentaje parcial explicado por los primeros m componentes principales, entonces,

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \mathrm{Var}(X_j)},$$

Con $m < p$

Generalmente no se escogen más de 3 componentes principales a fin de poderlas representar gráficamente, pero esto depende del porcentaje de variabilidad parcial recogido por las m componentes.

Toda la teoría expuesta anteriormente acerca del análisis de componentes principales, ha sido en base a una población, donde $\sum$ contiene las varianzas y covarianzas poblacionales. Sin embargo, el ejemplo dado en esta memoria con la base de datos FAMuSS se realiza a partir de datos muestrales, por lo cual se estima $\sum$ con S, la matriz de varianzas y covarianzas de la muestra.

2.4) Biplots

Dado un conjunto de datos multivariados, es posible realizar una representación gráfica de los mismos haciendo uso de un diagrama denominado biplot. En este tipo de diagrama, se utilizan las dos primeras componentes principales y es posible representar variables e individuos simultáneamente. Este diagrama es útil para realizar la interpretación de manera visual de las componentes calculadas, esto se hace de acuerdo a la dirección que tengan las variables que se representan mediante flechas. Un biplot se considera también una generalización multivariada de un diagrama bivariante.

Sea X la matriz de datos (según el método multivariado a utilizar, X puede ser de variables cuantitativas, una matriz de correlación, una tabla de contingencia, una matriz de coeficientes de regresión, etc.), para la construcción del biplot es necesario realizar una factorización de esta matriz, así:

$$X_{nxp} = F_{nxr} G'_{rxp}$$

Donde F y G se consideran matrices de marcadores fila y columna respectivamente, es decir, son los vectores cuyo producto interno da como resultado la mejor aproximación de la matriz dada X .

Esta factorización puede ser obtenida a partir de la descomposición en valores singulares (SVD) de la matriz X, pues esta descomposición garantiza la obtención de una aproximación de X del rango dado optimo en el sentido de mínimos cuadrados.

$$X = UDV' = FG'$$

Los marcadores de fila y columna varían según los valores singulares. De manera general éstos pueden ser:

$$F = UD^{\alpha} \text{ y } G = VD^{1-\alpha} \text{ donde } 0 < \alpha < 1$$

Para el caso de biplots en componentes principales se deben ajustar los marcadores teniendo en cuenta que si los valores de la matriz X se centran, entonces las componentes de la matriz $X'X$ son proporcionales a la matriz:

$$S = \frac{1}{n-1} X'X$$

Y podrían tomarse como marcadores:

$$F = \sqrt{n-1}U$$
$$G = \frac{1}{\sqrt{n-1}}VD$$

Al realizar un biplot con estas características, se conserva la métrica de las columnas, esto es, los productos escalares de los marcadores de las columnas son iguales a los productos escalares de las columnas de X, quienes a su vez son las varianzas y las covarianzas.

Capítulo 3. Aplicación del ACP

3) Análisis de componentes principales a la a la base de datos FAMuSS

3.1) Exploración de base de datos

Para realizar el análisis estadístico se emplea el software R. Primero se realiza la lectura de datos:

```
> str(X)
'data.frame':   1397 obs. of  347 variables:
 $ id                 : Factor w/ 1396 levels "FA-1801","FA-1802",..: 1 2 3 4 5 6 7 8…
 $ acdc_rs1501299     : Factor w/ 3 levels "AA","CA","CC": 2 2 2 3 2 3 3 3 3 3 ...
 $ ace_id             : Factor w/ 3 levels "DD","ID","II": 1 2 2 1 2 2 3 2 2 2 ...
 $ actn3_r577x        : Factor w/ 3 levels "CC","CT","TT": 1 2 2 2 1 2 3 2 2 1 ...
 $ actn3_rs540874     : Factor w/ 3 levels "AA","GA","GG": 3 2 2 2 3 2 1 2 2 3 ...
 $ actn3_rs1815739    : Factor w/ 3 levels "CC","TC","TT": 1 2 2 2 1 2 3 2 2 1 ...
 $ actn3_1671064      : Factor w/ 3 levels "AA","GA","GG": 1 2 2 2 1 2 3 2 2 1 ...
 $ ardb1_1801253      : Factor w/ 3 levels "CC","CG","GG": NA NA NA NA NA NA NA NA NA...
 $ adrb2_1042713      : Factor w/ 4 levels "AA","AG","GA",..: 3 3 3 1 3 3 3 3 4 4 ...
 $ adrb2_1042714      : Factor w/ 3 levels "CC","CG","GG": 2 1 2 1 2 2 2 2 3 3 ...
 $ adrb2_rs1042718    : Factor w/ 3 levels "AA","CA","CC": 3 2 3 3 3 3 3 3 3 3 ...
 $ adrb3_4994         : Factor w/ 3 levels "CC","TC","TT": 3 3 NA 3 3 2 3 NA 3 3 ...
```

Hay 1397 observaciones y 347 variables, de las cuales 225 corresponden a variables de tipo SNP. No todas las variables SNP tienen la misma cantidad de genotipos como se muestra a continuación:

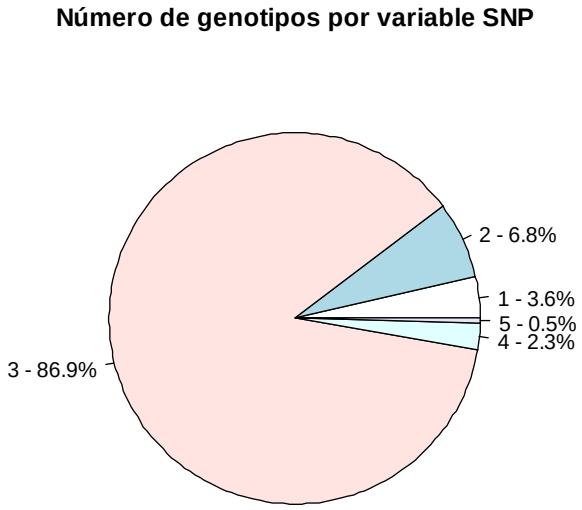

Figura 3.1Diagrama pastel del número de SNPs de acuerdo al número de genotipos

Para la Figura 3.1 se tomaron los 225 SNPs de la base de datos FAMuSS y se aprecian la cantidad de genotipos que aparecen en ellos. En su mayoría, el 86,9% de los

marcadores son bialelicos, y por lo tanto tienen 1, 2 o 3 genotipos. Aunque existen algunos pocos con tres alelos y por lo tanto puede haber hasta 6 genotipos y por esta razón hay un segmento 4 y 5 en el diagrama pastel.

Lo más habitual es encontrar polimorfismos con 1, 2 o 3 genotipos. Cuando solo hay un genotipo en el marcador indica que el SNP es monomórfico y por lo tanto todo el mundo tiene el mismo genotipo. Cuando hay 2, típicamente indica que existe solamente un homocigoto y que el otro homocigoto es ausente.

3.2) Análisis de las variables de la proteína alfa-actinina 3 ó actn3 (Alpha-actinin-3).

Las variables que se utilizan en primera instancia son 4 SNPs del gen de la musculatura actn3: r577x, rs540874, rs1815739 y 1671064. Estos SNPs se conocen por ser de los más estudiados en cuanto a los que afectan la estructura muscular. Codifica para una proteína de 901 aminoácidos y se encuentra situado en el cromosoma 11.

Realizando una exploración de estos datos, se encuentra inicialmente que el porcentaje de valores perdidos por cada una de estas variables es:

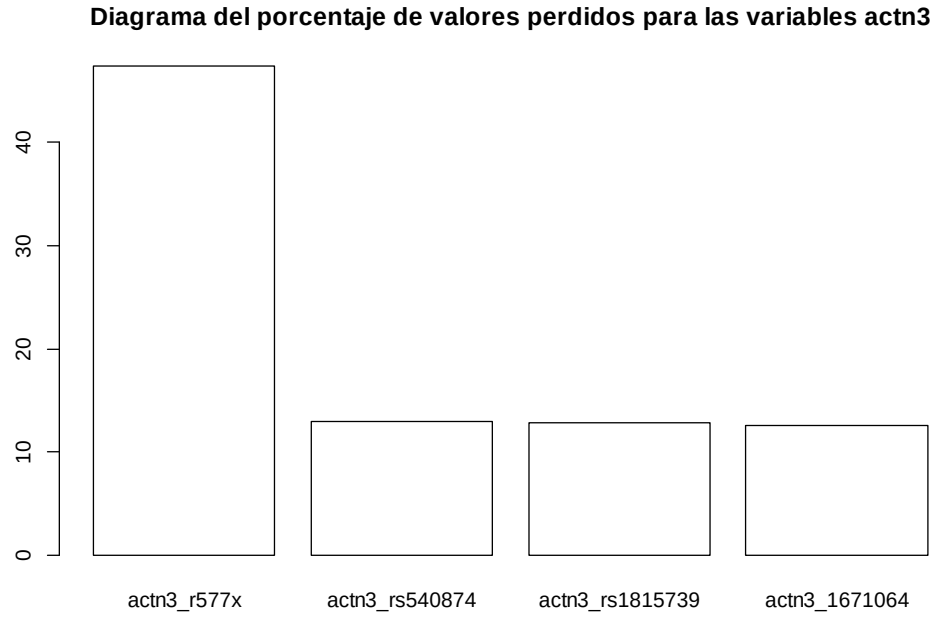

Figura 3.2 Porcentaje de valores perdidos de los SNPs actn3

Se observa que la variable actn3_r577x es la que presenta mayor porcentaje de datos faltantes con un 47.39% de información perdida. Las otras tres variables presentan a lo más, aproximadamente un 13% de valores missing.

Para efectos de simplicidad, se omiten los valores faltantes y se realiza el análisis estadístico sin tener en cuenta esta información.

3.2.1) Codificación de variables actn3 según el alelo menor

Una vez estudiados los casos de valores faltantes, se procede a realizar la codificación de las variables actn3 de acuerdo al alelo menor. Para ello se empieza calculando las tablas de frecuencias genotípicas y alélicas absolutas y relativas de cada una de las variables.

Frecuencias genotípicas:

actn3_r577x	Frecuencias Absolutas	Frecuencias Genotípicas
CC	216	0,2939
CT	318	0,4327
TT	201	0,2735
Total	**735**	**1,0000**

Tabla 3.1 Frecuencias genotípicas de la variable actn3_r577x

actn3_rs540874	Frecuencias Absolutas	Frecuencias Genotípicas
AA	226	0,1859
GA	595	0,4893
GG	395	0,3248
Total	**1216**	**1,0000**

Tabla 3.2 Frecuencias genotípicas de la variable actn3_rs540874

actn3_rs1815739	Frecuencias Absolutas	Frecuencias Genotípicas
CC	397	0,3262
TC	595	0,4889
TT	225	0,1849
Total	**1217**	**1,0000**

Tabla 3.3 Frecuencias genotípicas de la variable actn3_rs1815739

actn3_1671064	Frecuencias Absolutas	Frecuencias Genotípicas
AA	394	0,3227
GA	594	0,4865
GG	233	0,1908
Total	**1221**	**1,0000**

Tabla 3.4 Frecuencias genotípicas de la variable actn3_1671064

Frecuencias alélicas:

- Variable actn3_r577x

$$f_C = \frac{2f_{CC} + f_{CT}}{2} = 0{,}5102$$

$$f_T = \frac{2f_{TT} + f_{CT}}{2} = 0{,}4898$$

Debido a que T tiene la frecuencia más baja, entonces T es el alelo menor y por lo tanto la codificación que se realiza es 2 para TT, 1 para CT y 0 para CC.

- Variable actn3_rs540874

$$f_A = \frac{2f_{AA} + f_{GA}}{2} = 0{,}4305$$

$$f_G = \frac{2f_{GG} + f_{GA}}{2} = 0{,}5695$$

Debido a que A tiene la frecuencia más baja, entonces A es el alelo menor y por lo tanto la codificación que se realiza es 2 para AA, 1 para GA y 0 para GG.

- Variable actn3_rs1815739

$$f_C = \frac{2f_{CC} + f_{CT}}{2} = 0{,}5707$$

$$f_T = \frac{2f_{TT} + f_{CT}}{2} = 0{,}4293$$

Debido a que T tiene la frecuencia más baja, entonces T es el alelo menor y por lo tanto la codificación que se realiza es 2 para TT, 1 para CT y 0 para CC.

- Variable actn3_1671064

$$f_A = \frac{2f_{AA} + f_{GA}}{2} = 0{,}5659$$

$$f_G = \frac{2f_{GG} + f_{GA}}{2} = 0{,}4341$$

Debido a que G tiene la frecuencia más baja, entonces G es el alelo menor y por lo tanto la codificación que se realiza es 2 para GG, 1 para GA y 0 para AA.

Utilizando el cálculo de los alelos menores, se procede a realizar la codificación de las variables, en conclusión se tiene que:

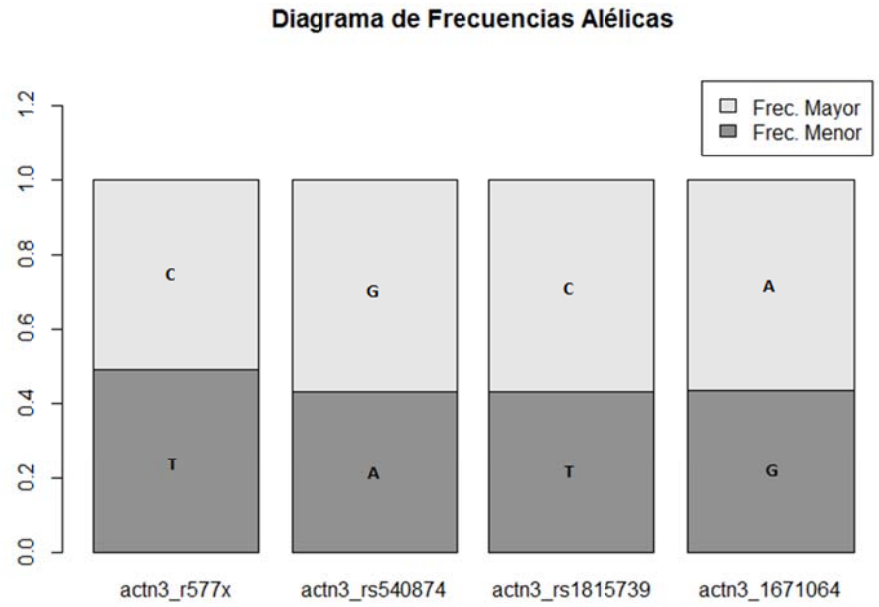

Figura 3.3 Frecuencias alélicas mayor y menor de r577x, rs540874, rs1815739 y 1671064

Para la variable:

- actn3_r577x el alelo menor es T.
- actn3_rs540874 el alelo menor es A.
- actn3_rs1815739 el alelo menor es T.
- actn3_1671064 el alelo menor es G.

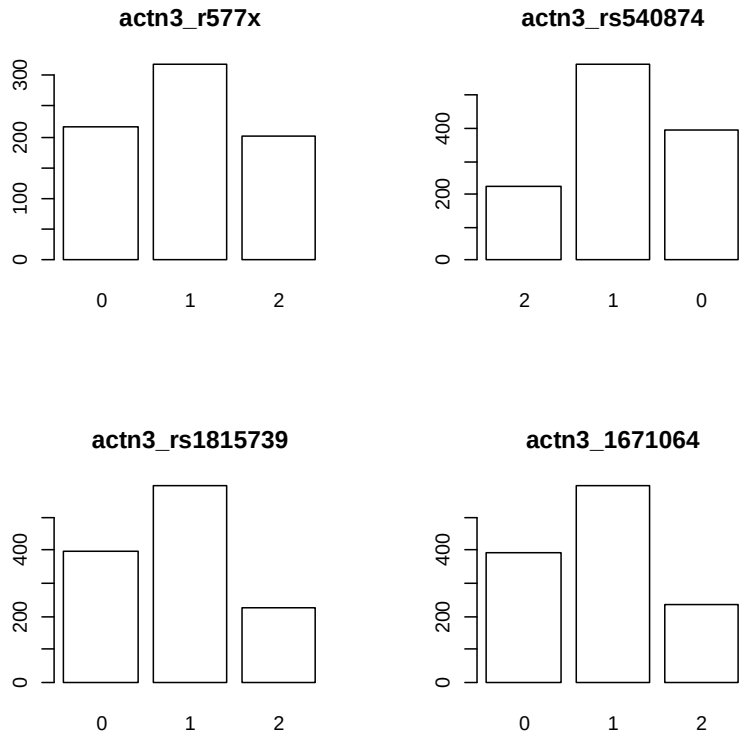

Figura 3.4 Diagrama de barras para las variables actn3 codificadas según el alelo menor

3.2.2) Análisis de componentes principales de los SNPs de actn3 basado en la matriz de correlaciones con la codificación del alelo menor

Para realizar el análisis de componentes principales se hace primero una exploración de las correlaciones con el propósito de saber si algunas de las componentes recogen parte de la variabilidad. Al calcular la matriz de correlaciones:

```
> Cor
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      1.0000000      0.7293332       0.7729220     0.7486462
actn3_rs540874   0.7293332      1.0000000       0.9574384     0.9659942
actn3_rs1815739  0.7729220      0.9574384       1.0000000     0.9775523
actn3_1671064    0.7486462      0.9659942       0.9775523     1.0000000
```

Se evidencia que las variables están altamente correlacionadas, por lo cual sería útil tomar en pocas componentes la mayor cantidad de información posible con el fin de quitar la variabilidad compartida.

En la matriz de varianzas y covarianzas:

```
> cov(Y)
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      0.5668065      0.3823311       0.4071816     0.3935795
actn3_rs540874   0.3823311      0.4848334       0.4664898     0.4696878
actn3_rs1815739  0.4071816      0.4664898       0.4896323     0.4776541
actn3_1671064    0.3935795      0.4696878       0.4776541     0.4876149
```

Se encuentra que los 4 SNPs tienen una varianza similar.

Se calculan los componentes principales basados en la matriz de correlaciones y se observa que la primera componente tiene varianza superior a 1, mientras que las demás no sobrepasan el 0,35.

```
> acp

Standard deviations:
[1] 1.8939245 0.5910183 0.2071031 0.1444157

Rotation:
           PC1        PC2         PC3         PC4
[1,] 0.4489352 -0.8902739 -0.06758798  0.03607434
[2,] 0.5116955  0.3113220 -0.78537716 -0.15629798
[3,] 0.5189027  0.1949562  0.54119920 -0.63232551
[4,] 0.5170734  0.2692267  0.29277638  0.75791426
```

Como se aprecia en el resultado anterior, la primera componente tiene varianza superior a 1. En las tres componentes restantes se presentan valores inferiores a uno, con descensos significativos.

```
> summary(acp)

Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.8939 0.59102 0.20710 0.14442
Proportion of Variance 0.8967 0.08733 0.01072 0.00521
Cumulative Proportion  0.8967 0.98406 0.99479 1.00000



> acp<-princomp(Y, scale=T, cor=T)
> acp

Call:
princomp(x = Y, cor = T, scale = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4
1.8939245 0.5910183 0.2071031 0.1444157

 4  variables and  724 observations.

> Tabla_de_Variabilidad_Procentual

                  Valores Propios    Varianza Explicada   Varianza Acumulada
Comp.1            3.58694983         0.896737457          0.8967375
Comp.2            0.34930260         0.087325651          0.9840631
Comp.3            0.04289168         0.010722921          0.9947860
Comp.4            0.02085589         0.005213972          1.0000000
```

Con la tabla de variabilidad se puede ver que con dos componentes principales se explica un 98.40% de la información original, lo cual se verifica claramente en la figura 3.5.

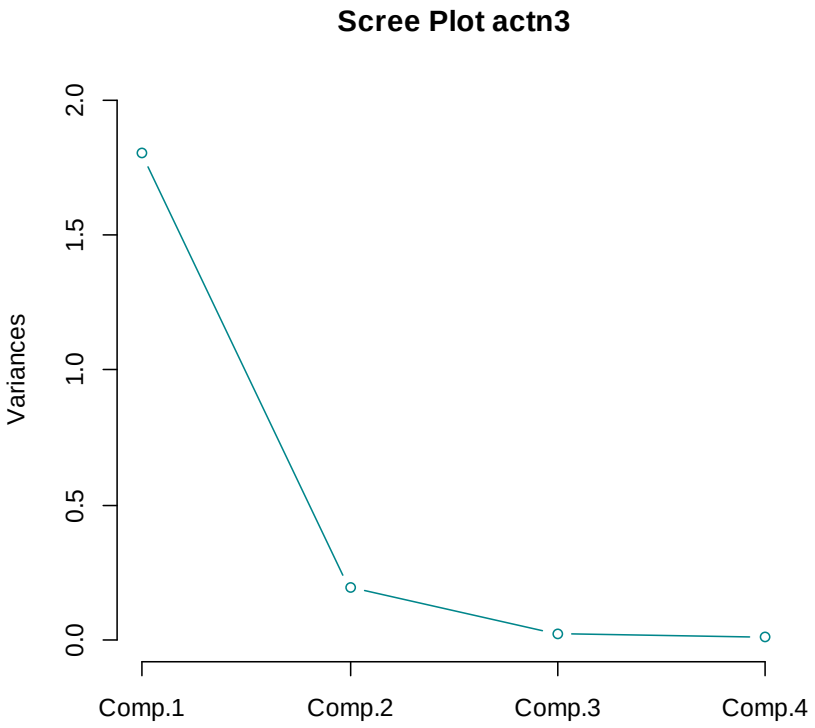

Figura 3.5 Scree Plot de las componentes principales

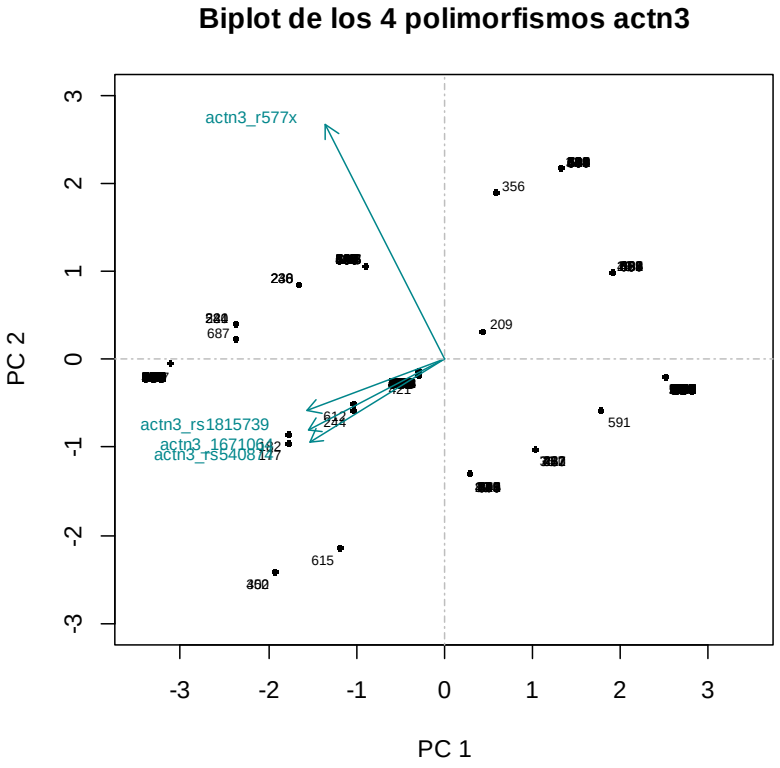

Figura 3.6 Biplot para los cuatro polimorfismos de actn3 con la codificación del alelo menor y utilizando la matriz de correlaciones.

En la figura 3.6 se observan 3 franjas de puntos alineados. La franja inferior representa los individuos que tienen cero copias del alelo menor para el marcador actn3_577x. Los puntos de la franja del medio corresponden a los individuos que tienen un alelo menor, son los heterocigotos. Los puntos de la franja superior corresponden a los individuos con dos alelos para este marcador.

Los puntos del lado izquierdo tienen 2 alelos para los 3 marcadores que apuntan hacia la izquierda y los que están a la izquierda tienen cero alelos para estos marcadores que apuntan a la izquierda. Es decir, el primer componente principal tiene una clara interpretación: separa los homocigotos del alelo menor de los homocigotos del alelo mayor.

Tres SNP que tienen las flechas casi coincidentes de la izquierda son tres marcadores con muchísima correlación y el otro marcador que apunta en la otra dirección es otra dimensión que no guarda tanta correlación con los tres que ya hay. Esta observación se puede verificar en la matriz de correlaciones, donde todas las correlaciones con 577x son relativamente más bajas.

3.2.3) Análisis de componentes principales de los SNPs de actn3 basado en la matriz de covarianzas con la codificación del alelo menor

Es de interés verificar los resultados obtenidos al calcular el análisis de componentes principales al utilizar no la matriz de correlaciones, si no la matriz de covarianzas como se muestra a continuación.

```
> ACP<-prcomp(Y, cor=F)
> ACP
Standard deviations:
[1] 1.3440332 0.4374882 0.1444964 0.1009296

Rotation:
                      PC1        PC2         PC3         PC4
actn3_r577x     0.4826718 -0.8730035 -0.06162606  0.03309294
actn3_rs540874  0.4998481  0.3258754 -0.78663513 -0.15862636
actn3_rs1815739 0.5101835  0.2198971  0.54236724 -0.63023476
actn3_1671064   0.5068458  0.2886442  0.28852241  0.75930671

> FP <- predict(ACP)
> summary(ACP)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.3440 0.43749 0.14450 0.10093
Proportion of Variance 0.8903 0.09434 0.01029 0.00502
Cumulative Proportion  0.8903 0.98469 0.99498 1.00000

> acp<-princomp(Y, cor=F)
```

```
> acp
Call:
princomp(x = Y, cor = F)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4
1.3431047 0.4371859 0.1443966 0.1008598

 4  variables and  724 observations.
```

Se logra determinar que al igual que el ACP con la matriz de correlaciones, el ACP con la matriz de covarianzas indica que con 2 componentes principales se recoge más del 98% de la información. Sin embargo, aunque la diferencia es muy poca se podría decir que para este caso en específico se ajusta mejor un análisis utilizando las covarianzas.

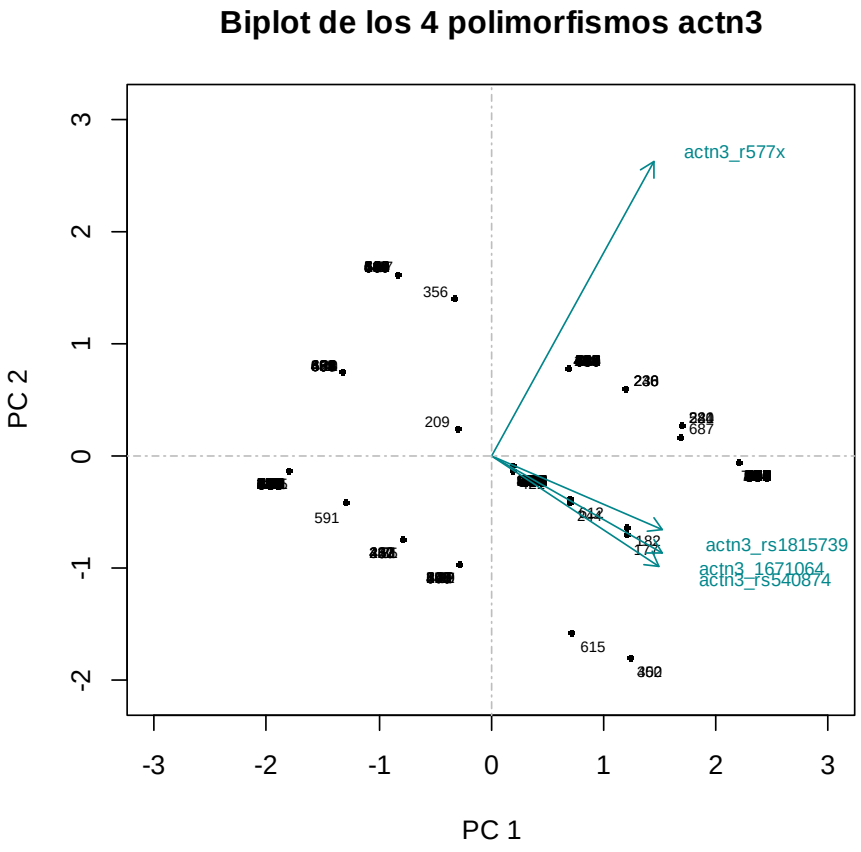

Figura 3.7 Biplot para los cuatro polimorfismos de actn3 con la codificación del alelo menor y utilizando la matriz de covarianzas

La figura 3.7 muestra que realmente no existe una diferencia muy significativa con respecto al análisis anterior. El biplot representado en la Figura 3.7 es un mero reflejo en el eje vertical del biplot anterior en la Figura 3.6.

3.2.4) Análisis de componentes principales de los SNPs de actn3 con el cambio de codificación según el alelo mayor

En este apartado se toman las mismas variables del gen actn3 y se realiza el análisis de componentes principales utilizando una codificación con respecto al alelo mayor. Se calculan la matriz de correlaciones y la matriz de varianzas y covarianzas.

```
> Cor
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      1.0000000      0.7293332       0.7729220     0.7486462
actn3_rs540874   0.7293332      1.0000000       0.9574384     0.9659942
actn3_rs1815739  0.7729220      0.9574384       1.0000000     0.9775523
actn3_1671064    0.7486462      0.9659942       0.9775523     1.0000000

> cov(Y)
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      0.5668065      0.3823311       0.4071816     0.3935795
actn3_rs540874   0.3823311      0.4848334       0.4664898     0.4696878
actn3_rs1815739  0.4071816      0.4664898       0.4896323     0.4776541
actn3_1671064    0.3935795      0.4696878       0.4776541     0.4876149
```

Se puede verificar que los resultados son invariantes con respecto a la codificación.

Para generalizar estos resultados, se demuestran las siguientes igualdades.

Sea x una variable codificada de acuerdo al alelo menor y y otra variable codificada respecto al alelo mayor. Es claro que la relación que presentan x y y es que $y = 2 - x$. A partir de esta definición se pretende demostrar que:

1. $var(x) = var(y)$

   Por definición de varianza se tiene que:

   $$var(x) = E(x - E[x])^2 = E(x^2) - [E(x)]^2$$

   De modo que al calcular la varianza de y se obtiene que:

   $$var(y) = var\ (2 - x) = \ E(2 - x)^2 - \{E(2 - x)\}^2$$

   $$= E\ (4 - 2x + x^2) - \{E(2) - E(x)\}^2$$

   $$= E(4) - 2E(x) + E(x^2) - \{E(2)\}^2 + E(2)E(x) - E(x)^2$$

   $$= 4 - 2E(x) + E(x^2) - 4 + 2E(x) - E(x)^2$$

   $$= E(x^2) - E(x)^2 = var(x)$$

2. $cov\ (x_1, x_2) = \ cov\ (y_1, y_2)$

   Por definición de covarianza se tiene que:

   $$cov\ (x_1, x_2) = E(x_1 x_2) - E(x_1)E(\ x_2)$$

   De modo que al calcular la covarianza de y_1 y y_2 se tiene que:

   $$si\ y_1 = 2 - x_1 \qquad y \qquad y_2 = 2 - x_2$$

$$= cov\,(y_1, y_2) = E(y_1 y_2) - E(y_1)E(y_2)$$

$$= E\{(2 - x_1)(2 - x_2)\} - E\{(2 - x_1)\}E\{(2 - x_2)\}$$

$$= E(4 - 2x_2 - 2x_1 + x_1 x_2) - \{E(2) - E(x_1)\}\{E(2) - E(x_2)\}$$

$$= E(4 - 2x_2 - 2x_1 + x_1 x_2) - [\{E(2)\}^2 - E(2)E(x_2) - E(2)E(x_1) + E(x_1)E(x_2)]$$

$$= E(4) - E(2x_2) - E(2x_1) + E(\,x_1 x_2) - \{E(2)\}^2 + E(2)E(x_2) + E(2)E(x_1) - E(x_1)E(x_2)$$

$$= 4 - 2E(x_2) - 2E(x_1) + E(x_1 x_2) - 4 + 2E(x_2) + 2E(x_1) - E(x_1)E(x_2)$$

$$= E(x_1 x_2) - E(x_1)E(x_2) = cov\,(x_1, x_2)$$

Al realizar el biplot para las variables del gen actn3 con respecto al cambio de codificación, se observa en la Figura 3.8, que el biplot obtenido es el mismo de la Figura 3.6.

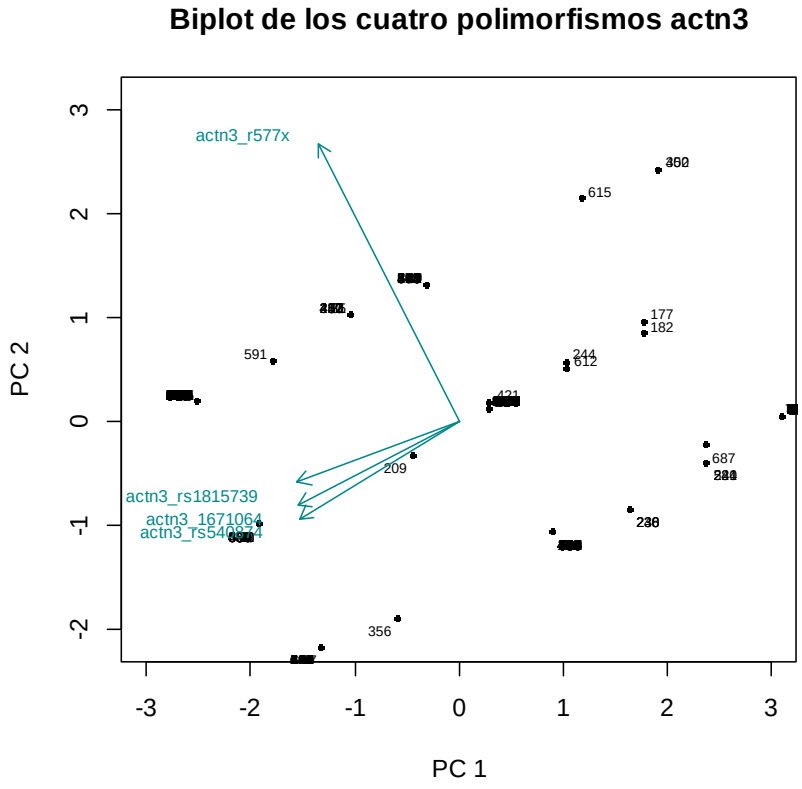

Figura 3.8 Biplot de los polimorfismos del gen actn3 con respecto a la codificación del alelo mayor

Indistinto de la codificación o del tipo de análisis de componentes principales, ya sea basado en el de la matriz de correlaciones o de covarianzas, para todos los casos considerando las variables del gen actn3, se puede considerar suficiente tomar únicamente la primera componente principal para el análisis de estas variables de SNPs. Aunque es importante notar, que el mejor ajuste esta dado cuando se realiza el análisis basado en la matriz de correlaciones.

3.2.5) Análisis de componentes principales de los SNPs del gen actn3 según la raza Caucasian

Es conocido que las frecuencias alélicas pueden variar de una población a otra. La base de datos de FAMuSS consta de individuos de raza Caucasiana y Afro-Americana. Se repite aquí el ACP, estratificando el análisis según la procedencia de los individuos.

Inicialmente se calcula una tabla de frecuencias para la variable raza:

Raza	Frecuencia Absoluta	Porcentaje
African Am	44	3,15
Am Indian	1	0,07
Asian	97	6,94
Caucasian	791	56,62
Hispanic	52	3,72
Other	49	3,51
NA's	363	25,98
Total	**1397**	**100,00**

Tabla 3.5 Tabla de frecuencias de la variable raza

Se puede notar, según la tabla 3.5, que el 52,62% de la información de la base de datos corresponde a individuos de la raza Caucasian. Mientras que tan solo el 3,15% es de la raza African American.

Se procede ahora a realizar el análisis de componentes principales para los individuos de la raza Caucasian, tomando los SNPs del gen actn3. Al calcular las matrices de covarianzas y de correlaciones se encuentra:

Matriz de covarianzas:

```
> cov(Y)
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      0.5400666      0.3550433       0.3676838     0.3592361
actn3_rs540874   0.3550433      0.4619765       0.4520409     0.4481279
actn3_rs1815739  0.3676838      0.4520409       0.4631845     0.4550433
actn3_1671064    0.3592361      0.4481279       0.4550433     0.4616656
```

Matriz de correlaciones:

```
> Cor
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      1.0000000      0.7108005       0.7351463     0.7194366
actn3_rs540874   0.7108005      1.0000000       0.9772164     0.9703498
actn3_rs1815739  0.7351463      0.9772164       1.0000000     0.9840382
actn3_1671064    0.7194366      0.9703498       0.9840382     1.0000000
```

Componentes principales a partir de la matriz de correlaciones:

```
> ACP<-prcomp(Y, scale=T, cor=T)
> ACP
Standard deviations:
[1] 1.8878641 0.6252139 0.1747471 0.1205826

Rotation:
```

```
                       PC1        PC2         PC3         PC4
actn3_r577x      0.4382812 -0.8982054  0.01893652 -0.02789264
actn3_rs540874   0.5165750  0.2752885  0.78102377 -0.21764273
actn3_rs1815739 0.5215930  0.2254802 -0.20218965  0.79763325
actn3_1671064    0.5186527  0.2580735 -0.59056128 -0.56181388

> FP <- predict(ACP)
> summary(ACP)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.888 0.62521 0.17475 0.12058
Proportion of Variance 0.891 0.09772 0.00763 0.00364
Cumulative Proportion  0.891 0.98873 0.99636 1.00000

> acp<-princomp(Y, scale=T, cor=T)
> acp
Call:
princomp(x = Y, cor = T, scale = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4
1.8878641 0.6252139 0.1747471 0.1205826

 4  variables and  475 observations.
```

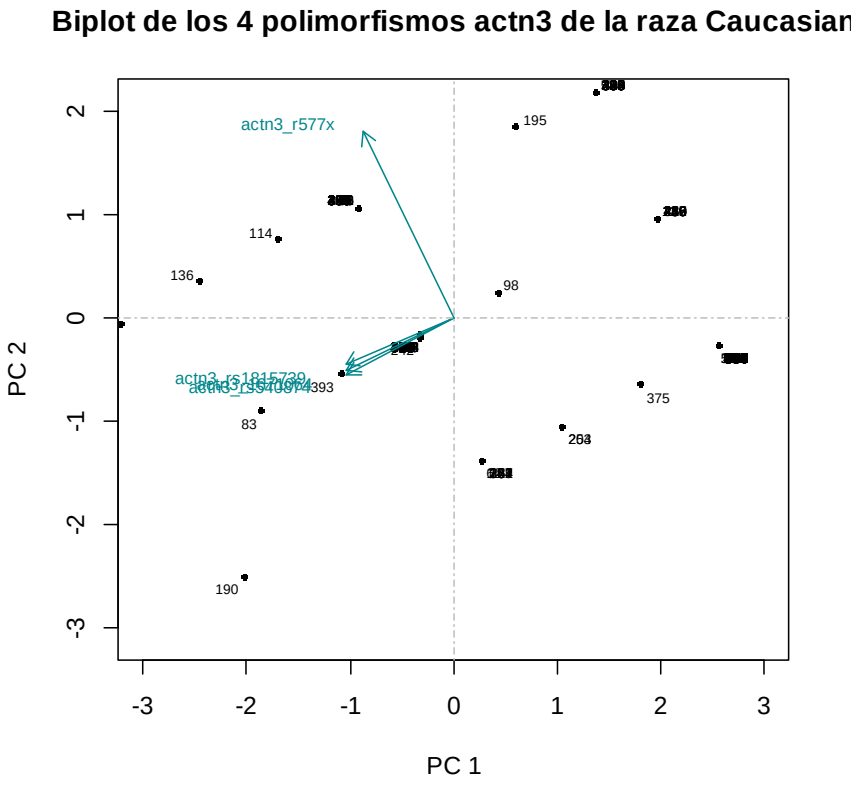

Figura 3.9 Biplot de los polimorfismos del gen actn3de la raza Caucasian con la matriz de correlaciones

En la Figura 3.9 se confirma lo calculado con la matriz de correlaciones, pues existen 3 variables del gen actn3 que se encuentran altamente correlacionadas: actn3_rs540874, actn3_rs1815739 y actn3_1671064. Se conserva la agrupación de individuos de acuerdo al conteo de alelos.

Componentes principales a partir de la matriz de covarianzas:

```
> ACP<-prcomp(Y, cor=F)
```

```
> ACP
Standard deviations:
[1] 1.30449513 0.45205293 0.11876910 0.08202251

Rotation:
                       PC1        PC2         PC3         PC4
actn3_r577x     0.4736353 -0.8801754  0.01737207 -0.02567108
actn3_rs540874  0.5058206  0.2939499  0.78134571 -0.21734262
actn3_rs1815739 0.5118007  0.2481549 -0.20297368  0.79704507
actn3_1671064   0.5078140  0.2780356 -0.58991421 -0.56286973

> FP <- predict(ACP)
> summary(ACP)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.3045 0.4521 0.11877 0.08202
Proportion of Variance 0.8831 0.1061 0.00732 0.00349
Cumulative Proportion  0.8831 0.9892 0.99651 1.00000

> acp<-princomp(Y, cor=F)
> acp
Call:
princomp(x = Y, cor = F)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4
1.30312125 0.45157684 0.11864401 0.08193612

 4    variables and  475 observations.
```

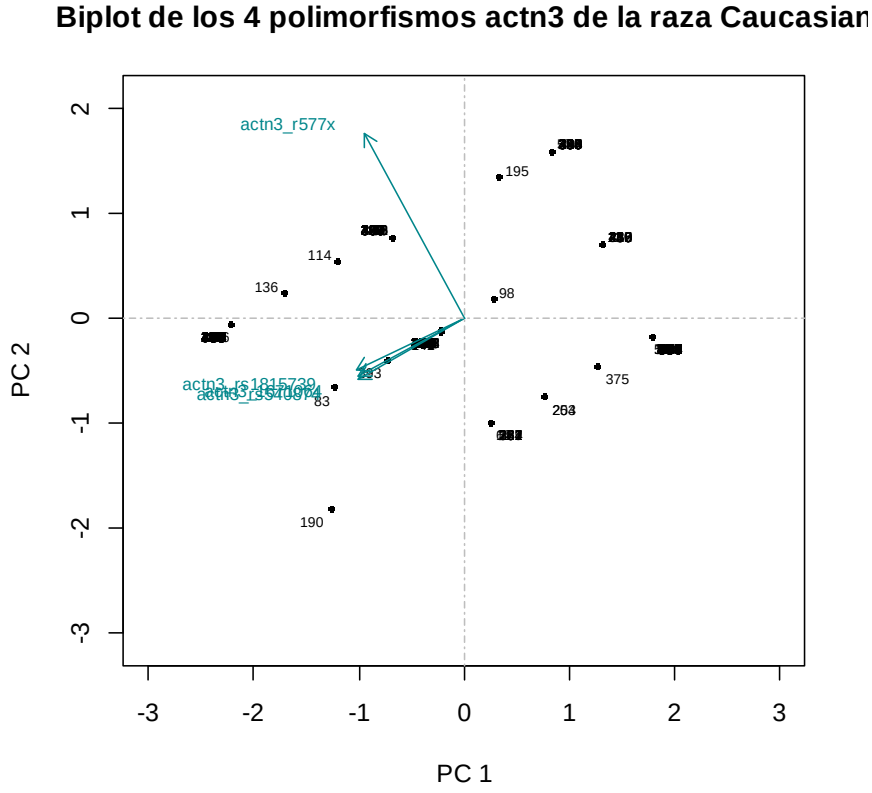

Figura 3.10 Biplot de los polimorfismos del gen actn3 de la raza Caucasian con la matriz de covarianzas

Se puede evidenciar en la Figura 3.10, que el análisis realizado a los individuos de la raza Caucasian se asemeja mucho a los resultados obtenidos para el análisis global. Esto

puede deberse a que el mayor porcentaje de información corresponde a individuos de esta raza.

3.2.6) Análisis de componentes principales de los SNPs del gen actn3 según la raza African American

En este apartado se muestran los resultados del análisis de componentes principales únicamente para los individuos de la raza African American

Matriz de covarianzas:

```
> cov(Y)
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      0.6243386      0.2328042       0.3015873     0.2328042
actn3_rs540874   0.2328042      0.5436508       0.3968254     0.5436508
actn3_rs1815739  0.3015873      0.3968254       0.4021164     0.3968254
actn3_1671064    0.2328042      0.5436508       0.3968254     0.5436508
```

Matriz de correlaciones:

```
> Cor
               actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064
actn3_r577x      1.0000000      0.3995957       0.6019038     0.3995957
actn3_rs540874   0.3995957      1.0000000       0.8487183     1.0000000
actn3_rs1815739  0.6019038      0.8487183       1.0000000     0.8487183
actn3_1671064    0.3995957      1.0000000       0.8487183     1.0000000
```

Todas las variables de actn3 para los individuos de la raza Afro-Americana se encuentran correlacionadas positivamente. Sin embargo, las variables actn3_1671064 y actn3_rs540874 presentan una correlación perfecta.

Componentes principales a partir de la matriz de correlaciones:

```
> ACP<-prcomp(Y, scale=T, cor=T)
> ACP
Standard deviations:
[1] 1.763132e+00 8.653022e-01 3.776469e-01 5.681312e-16

Rotation:
                      PC1         PC2        PC3           PC4
actn3_r577x     -0.3579962  0.88734331 -0.2906210  0.000000e+00
actn3_rs540874  -0.5401299 -0.32337646 -0.3220052 -7.071068e-01
actn3_rs1815739 -0.5369898  0.05896764  0.8415252  1.332268e-15
actn3_1671064   -0.5401299 -0.32337646 -0.3220052  7.071068e-01

> FP <- predict(ACP)
> summary(ACP)
Importance of components:
                          PC1    PC2     PC3       PC4
Standard deviation     1.7631 0.8653 0.37765 5.681e-16
Proportion of Variance 0.7772 0.1872 0.03565 0.000e+00
Cumulative Proportion  0.7772 0.9644 1.00000 1.000e+00

> acp<-princomp(Y, scale=T, cor=T)
> acp
Call:
princomp(x = Y, cor = T, scale = T)

Standard deviations:
```

```
      Comp.1       Comp.2       Comp.3       Comp.4
1.763132e+00 8.653022e-01 3.776469e-01 1.290478e-08

 4  variables and  28 observations.
```

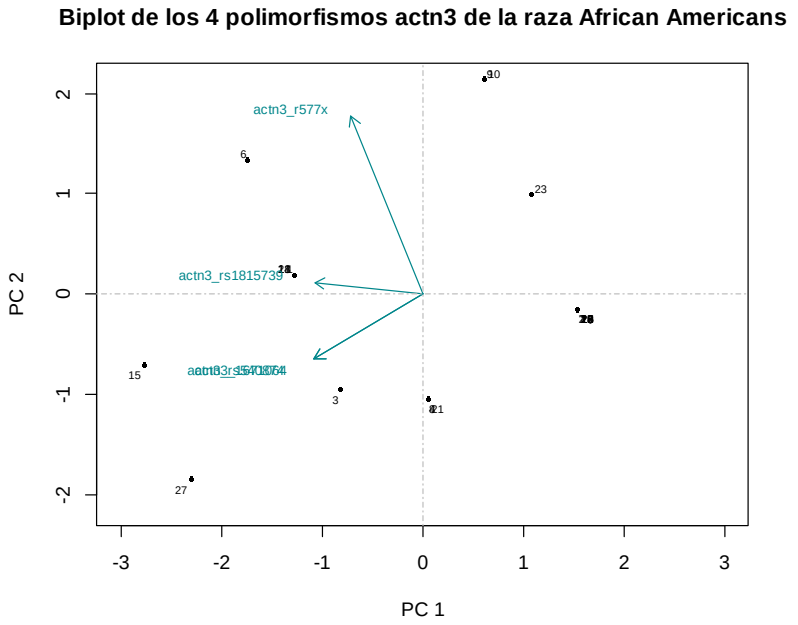

Figura 3.11 Biplot de los polimorfismos del gen actn3de la raza African Americans con la matriz de correlaciones

En la Figura 3.11 se puede notar que únicamente aparecen 3 flechas, habiendo 4 variables, esto indica que dos de las variables (actn3_1671064 y actn3_rs540874) tienen una dependencia total, lo cual ya se había notado en la matriz de correlaciones.

Componentes principales a partir de la matriz de covarianzas:

```
> ACP<-prcomp(Y, cor=F)
> ACP
Standard deviations:
[1] 1.264012e+00 6.734235e-01 2.500605e-01 1.188343e-16

Rotation:
                       PC1         PC2        PC3           PC4
actn3_r577x     -0.4100295  0.88136249 -0.2346827  0.000000e+00
actn3_rs540874  -0.5526039 -0.33386989 -0.2883745 -7.071068e-01
actn3_rs1815739 -0.4702485  0.01618499  0.8823856  5.551115e-17
actn3_1671064   -0.5526039 -0.33386989 -0.2883745  7.071068e-01

> FP <- predict(ACP)
> summary(ACP)
Importance of components:
                          PC1    PC2     PC3       PC4
Standard deviation     1.2640 0.6734 0.25006 1.188e-16
Proportion of Variance 0.7559 0.2145 0.02958 0.000e+00
Cumulative Proportion  0.7559 0.9704 1.00000 1.000e+00

> acp<-princomp(Y, cor=F)
> acp
Call:
princomp(x = Y, cor = F)

Standard deviations:
```

```
      Comp.1       Comp.2       Comp.3       Comp.4
1.241235e+00 6.612887e-01 2.455545e-01 1.249500e-08

 4  variables and  28 observations.
```

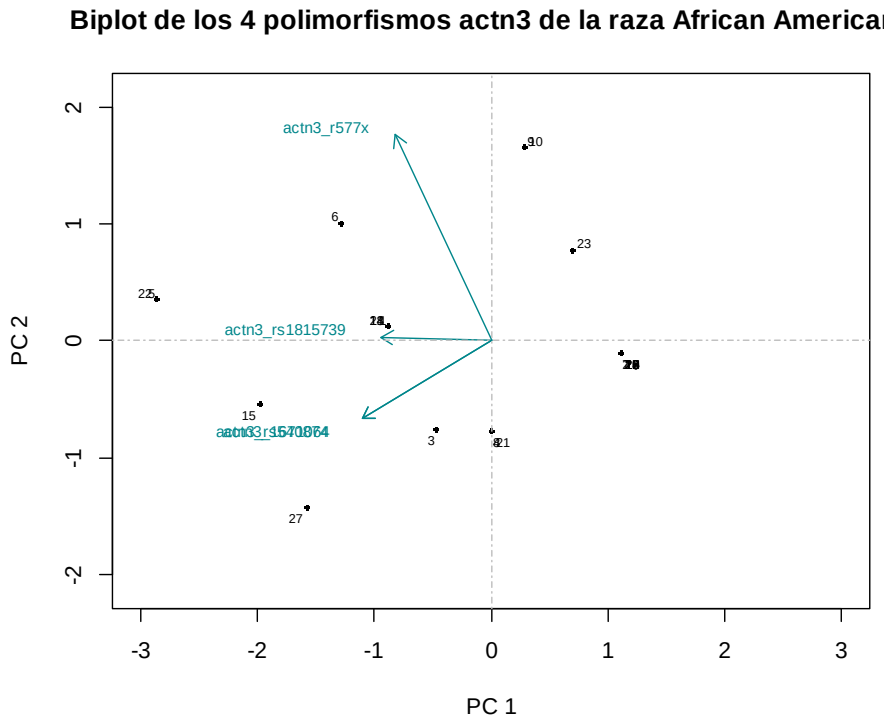

Figura 3.12 Biplot de los polimorfismos del gen actn3 de la raza African Americans con la matriz de covarianzas

Los individuos de la raza African American presentan un comportamiento diferente en cuanto al gen actn3, pues existe un SNP que coincide con la primera componente principal y aparentemente este es diferente para caucasianos y africanos.

3.3) Análisis de las variables Akt1

En este apartado se aborda el ACP de un conjunto más grande de SNPs, los SNPs que forman parte del gen Akt1.

Las varianzas para las 28 variables Akt1 se muestran a continuación:

```
> Var
       akt1_t22932c         akt1_g15129a         akt1_g14803t
          0.07502750           0.46809681           0.47326733

akt1_c10744t_c12886t akt1_t10726c_t12868c akt1_t10598a_t12740a
          0.02909180           0.18662755           0.49623518

 akt1_c9756a_c11898t          akt1_t8407g          akt1_a7699g
          0.42769833           0.18252047           0.16480870

  akt1_c6148t_c8290t   akt1_c6024t_c8166t   akt1_c5854t_c7996t
          0.18809436           0.43158538           0.22222222

   akt1_c832g_c3359g           akt1_g288c    akt1_g1780a_g363a
          0.13429899           0.48585747           0.02178218
```

```
akt1_g2347t_g205t    akt1_g2375a_g233a          akt1_g4362c
       0.47503973           0.24548344           0.15820804

      akt1_c15676t         akt1_a15756t         akt1_g20703a
       0.26644664           0.42058428           0.21332355

      akt1_g22187a         akt1_a22889g         akt1_g23477a
       0.37436744           0.48725095           0.45951595

      akt2_7254617        akt2_rs892118         akt2_2304186
       0.28105366           0.27826672           0.51592715

       akt2_969531
       0.41386139
```

Al realizar el ACP de este grupo de polimorfismos, se encuentra:

```
> summary(acp2)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.8238 2.3466 1.61924 1.52103 1.41442 1.23814 1.00901
Proportion of Variance 0.2848 0.1967 0.09364 0.08263 0.07145 0.05475 0.03636
Cumulative Proportion  0.2848 0.4814 0.57509 0.65771 0.72916 0.78391 0.82027
                          PC8    PC9   PC10   PC11    PC12    PC13    PC14
Standard deviation     0.9928 0.9286 0.7274 0.6715 0.59825 0.50530 0.43852
Proportion of Variance 0.0352 0.0308 0.0189 0.0161 0.01278 0.00912 0.00687
Cumulative Proportion  0.8555 0.8863 0.9052 0.9213 0.93405 0.94317 0.95004
                         PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.4300 0.42120 0.39970 0.38422 0.37862 0.34991 0.34324
Proportion of Variance 0.0066 0.00634 0.00571 0.00527 0.00512 0.00437 0.00421
Cumulative Proportion  0.9566 0.96298 0.96868 0.97396 0.97908 0.98345 0.98766
                          PC22   PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.33198 0.2900 0.22412 0.20024 0.17143 0.14801 0.09840
Proportion of Variance 0.00394 0.0030 0.00179 0.00143 0.00105 0.00078 0.00035
Cumulative Proportion  0.99159 0.9946 0.99639 0.99782 0.99887 0.99965 1.00000
```

Como lo muestra el resultado anterior, las 7 primeras componentes tienen varianza superior a 1. Las siguientes varianzas presentan un descenso progresivo.

```
> ACP2<-princomp(Z, scale=T, cor=T)
> ACP2
Call:
princomp(x = Z, cor = T, scale = T)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
2.82383675 2.34657367 1.61924307 1.52102832 1.41441604 1.23813743 1.00901028
    Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14
0.99283169 0.92858879 0.72739683 0.67146999 0.59824513 0.50529780 0.43851697
   Comp.15    Comp.16    Comp.17    Comp.18    Comp.19    Comp.20    Comp.21
0.43002100 0.42119522 0.39969925 0.38422161 0.37862033 0.34991241 0.34323730
   Comp.22    Comp.23    Comp.24    Comp.25    Comp.26    Comp.27    Comp.28
0.33197705 0.28997791 0.22412240 0.20024405 0.17143122 0.14800598 0.09839579

 28  variables and  405 observations.
```

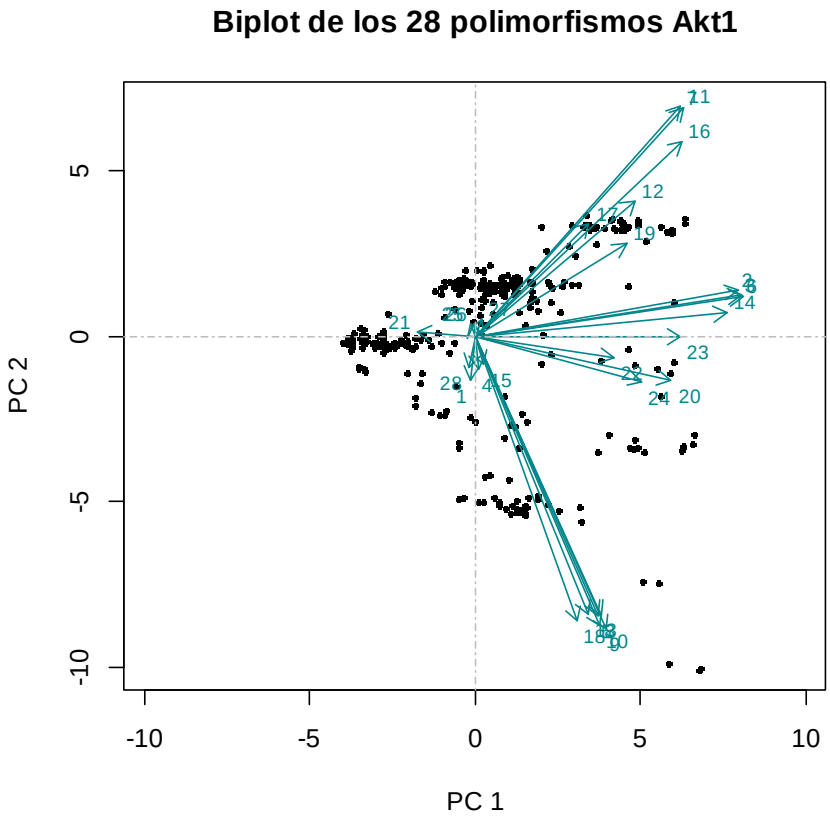

Figura 3.13 Biplot de los polimorfismos del gen akt1con la matriz de correlaciones

En la Figura 3.13 se muestra que existen variables del gen akt1 que presentan correlaciones altas. El biplot verifica que se detectan grupos de polimorfismos relacionados.

```
> Tabla_de_Variabilidad_Procentual
   Valores Propios Varianza Explicada Varianza Acumulada
1       7.974053992         0.2847876426          0.2847876
2       5.506407981         0.1966574279          0.4814451
3       2.621948114         0.0936410041          0.5750861
4       2.313527143         0.0826259694          0.6577120
5       2.000572738         0.0714490263          0.7291611
6       1.532984308         0.0547494396          0.7839105
7       1.018101735         0.0363607763          0.8202713
8       0.985714767         0.0352040988          0.8554754
9       0.862277138         0.0307956121          0.8862710
10      0.529106155         0.0188966484          0.9051676
11      0.450871944         0.0161025694          0.9212702
12      0.357897237         0.0127820442          0.9340523
13      0.255325864         0.0091187809          0.9431710
14      0.192297133         0.0068677548          0.9500388
15      0.184918065         0.0066042166          0.9566430
16      0.177405412         0.0063359076          0.9629789
17      0.159759491         0.0057056961          0.9686846
18      0.147626245         0.0052723659          0.9739570
19      0.143353352         0.0051197626          0.9790767
20      0.122438696         0.0043728106          0.9834496
21      0.117811842         0.0042075658          0.9876571
22      0.110208765         0.0039360273          0.9915931
23      0.084087187         0.0030031138          0.9945963
```

```
24     0.050230850       0.0017939589          0.9963902
25     0.040097680       0.0014320600          0.9978223
26     0.029388665       0.0010495952          0.9988719
27     0.021905770       0.0007823489          0.9996542
28     0.009681731       0.0003457761          1.0000000
```

Con la tabla de variabilidad se puede ver que con dos componentes principales se explica un 48% de la información original. Generalmente con dos componentes se puede explicar mucha más información, sin embargo con los datos de este ejemplo no y está claro que puede deberse al hecho que las correlaciones de las variables son relativamente bajas. Se evidencia entonces que para llegar a explicar más de 90% se necesitarían por lo menos 10 componentes principales. Se validan estos resultados con la Figura 3.14.

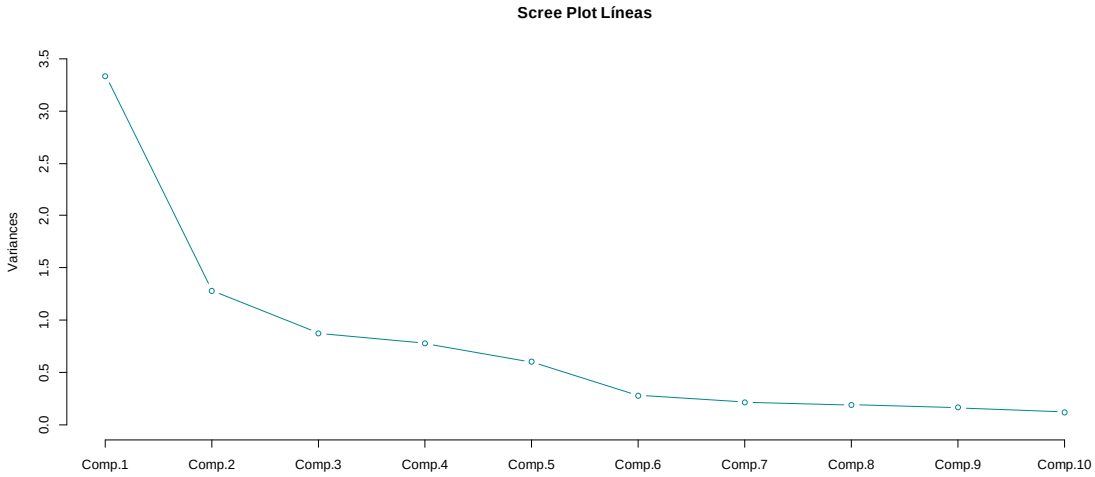

Figura 3.14 Scree Plot de los polimorfismos del gen Akt1

Componentes principales a partir de la matriz de covarianzas:

```
> summary(acp2)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.8280 1.1345 0.9363 0.88318 0.77694 0.53049 0.46543
Proportion of Variance 0.3896 0.1501 0.1022 0.09095 0.07038 0.03281 0.02526
Cumulative Proportion  0.3896 0.5397 0.6419 0.73284 0.80322 0.83603 0.86129
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.43748 0.40735 0.34723 0.32401 0.31735 0.28006 0.26449
Proportion of Variance 0.02232 0.01935 0.01406 0.01224 0.01174 0.00915 0.00816
Cumulative Proportion  0.88361 0.90295 0.91701 0.92925 0.94099 0.95014 0.95830
                          PC15    PC16    PC17   PC18   PC19    PC20    PC21
Standard deviation     0.25760 0.24783 0.21284 0.2071 0.1830 0.15972 0.14959
Proportion of Variance 0.00774 0.00716 0.00528 0.0050 0.0039 0.00297 0.00261
Cumulative Proportion  0.96603 0.97319 0.97848 0.9835 0.9874 0.99036 0.99296
                          PC22   PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.12227 0.1135 0.10411 0.09089 0.07809 0.06421 0.05647
Proportion of Variance 0.00174 0.0015 0.00126 0.00096 0.00071 0.00048 0.00037
Cumulative Proportion  0.99471 0.9962 0.99747 0.99844 0.99915 0.99963 1.00000

> ACP2<-princomp(Z, cor=F)
> ACP2
Call:
princomp(x = Z, cor = F)
```

```
Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
1.82570800 1.13312517 0.93510698 0.88208946 0.77597730 0.52983603 0.46485834
    Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14
0.43694267 0.40684309 0.34680257 0.32360713 0.31696045 0.27971879 0.26416662
   Comp.15    Comp.16    Comp.17    Comp.18    Comp.19    Comp.20    Comp.21
0.25728019 0.24752515 0.21257509 0.20684618 0.18274925 0.15952362 0.14940081
   Comp.22    Comp.23    Comp.24    Comp.25    Comp.26    Comp.27    Comp.28
0.12212021 0.11335299 0.10397929 0.09077867 0.07799672 0.06412912 0.05640062

 28  variables and  405 observations.
```

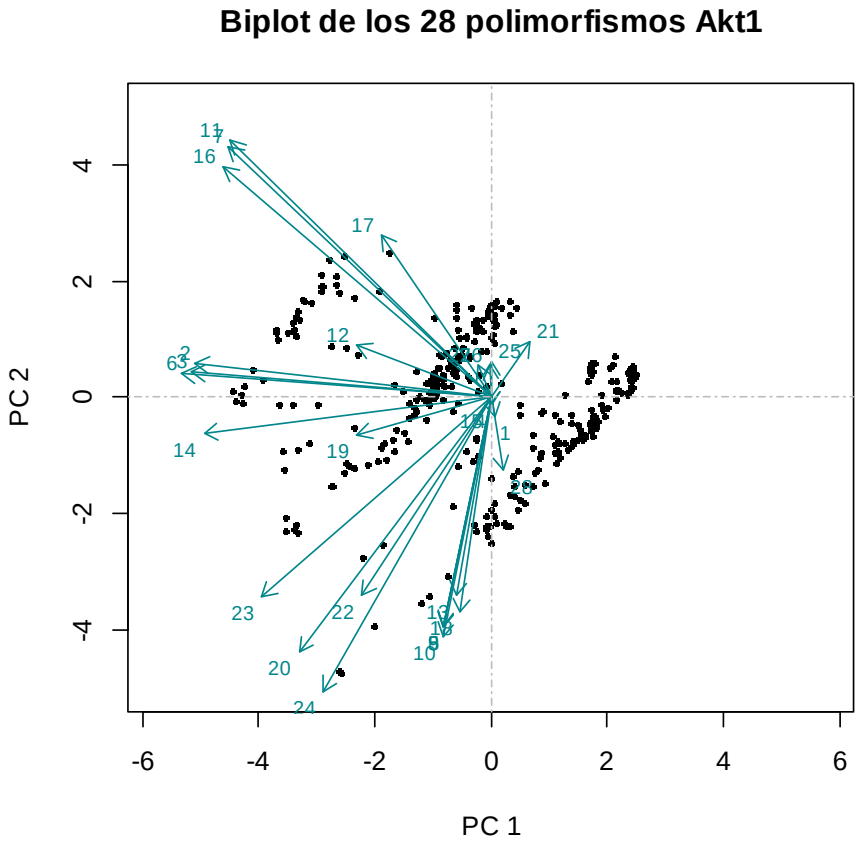

Figura 3.15 Biplot de los polimorfismos del gen akt1 con la matriz de covarianzas

Al realizar el análisis de componentes principales basados en la matriz de covarianzas, se encuentra que el ajuste es mejor, pues el biplot de la Figura 3.15 muestra tres franjas de puntos alineados que indican el número de alelos que presentan los individuos para este gen. La franja superior indica los individuos que no tienen copias del alelo menor en este marcador. Los de la franja central corresponden a los heterocigotos o los que presentan un alelo menor y los de la franja inferior son los que presentan dos alelos para este polimorfismo. Se visualiza además que existen variables correlacionadas entre si y que hay unos pocos polimorfismos separan estos tres grupos.

4) Conclusiones Generales

Debido a la trascendencia y la importancia que pueden tener los estudios relacionados con datos genéticos, se destacan los resultados aquí encontrados para continuar con la búsqueda de técnicas que permitan analizar de manera práctica y sencilla información de este tipo. Se muestra que el análisis de componentes principales es de gran utilidad pues permite realizar reducciones en la dimensión de la base de datos sin pérdida significativa de información.

Se hace necesaria una exploración inicial de los datos con el fin de tener una idea de la información a utilizar, realizando tablas de frecuencias y gráficos donde se pueda detallar la cantidad de datos faltantes por variable. Es importante tener en cuenta para estudios posteriores el manejo de los datos missing, pues podría arrojar resultados diferentes e interesantes.

Para estudios de datos genéticos que involucren variables de SNPs, se hace necesaria la codificación de los mismos debido a que los resultados iniciales vienen dados como una variable categórica producto de la combinación de los alelos que pueda presentar el gen a estudiar. Esta codificación se realiza con base en el conteo de los alelos y dependiendo del número de ellos (generalmente bialelicos), puede tomar los valores 0, 1 o 2.

Al aplicar el análisis de componentes principales en un grupo de datos del gen actn3 y realizar la codificación de las variables según el alelo menor y según el alelo mayor se encontró que no existen diferencias en los resultados, es decir, el ACP es invariante respecto a la codificación de los SNPs. Esto se generaliza a partir de la demostración matemática de la igualdad de varianzas y de covarianzas, donde se pudo encontrar que $var(x) = var(y)$ y $cov\,(x_1, x_2) = cov\,(y_1, y_2)$ donde x es una variable codificada de acuerdo al alelo menor y y otra variable codificada respecto al alelo mayor.

El biplot es una herramienta valiosa para la interpretación de los resultados. Se pudo evidenciar que siempre existían grupos de variables altamente correlacionadas y en la mayoría de los biplots se formaron 3 franjas con nubes de puntos alineados que indican la presencia o ausencia del alelo según el conteo, ya sea el alelo menor o el alelo mayor.

Los resultados obtenidos permiten determinar también, que el ACP sirve para analizar datos genéticos, específicamente datos de SNPs, donde a partir del biplot se pudo establecer que se presentan grupos de SNPs relacionados.

Sería interesante poder examinar más grupos de genes de SNPs presentados en la base de datos FAMuSS e incluso seguir buscando bases de datos que contengan información de polimorfismos que permitan seguir encontrando patrones por genes y según la raza de los individuos.

5) Bibliografía

Foulkes, A. S. (2009). *Applied Statistical Genetics with R For Population-based Association Studies.* USA: Springer.

Johnson & Wichern. (2002). *Applied Multivariate Statistical.* Prentice Hall.

Jolliffe, I. (1986). *Principal Component Analysis.* New York: Springer-Verlag.

Manolio, T. A. (2010). Genomewide Association Studies and Assessment of the Risk of Disease. *The New England Journal of Medicine 363*, 166.

Pearson, T. A. (2008). How to Interpret a Genome-wide Association Study. *The Journal of the American Medical Association 299*, 1335 - 1344.

Thompson, P. D. (2004). Functional Polymorphisms Associated with Human Muscle Size and Strength. *Medicine & Science In Sports & Exercise*, 1132 - 1139.

Warnes, G. (s.f.). *genetics. R-package version 1.3.8.1. URL http://CRAN.R-project.org/package=genetics*.